DOCUMENT RESUME

ED 361 386 TM 020 486

AUTHOR Hewitson, Mal
TITLE The Concept of Performance Levels in Criterion-Referenced Assessment.
PUB DATE [88]
NOTE 22p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Standards; *Achievement Tests; *Cost Effectiveness; *Criterion Referenced Tests; Diagnostic Tests; *Educational Assessment; Elementary Secondary Education; Foreign Countries; Mastery Learning; *Mastery Tests; Student Evaluation; *Testing Problems; Test Reliability; Test Validity
IDENTIFIERS Performance Based Evaluation; *Performance Levels; Standard Setting

ABSTRACT

The concept of performance levels in criterion-referenced assessment is explored by applying the idea to different types of tests commonly used in schools, mastery tests (including diagnostic tests) and achievement tests. In mastery tests, a threshold performance standard must be established for each criterion. Attainment of this threshold signifies mastery. Performance levels in diagnostic testing are a special category of performance levels in mastery tests, but the focus is on the learning function rather than the certification function of the tests. Achievement tests are designed to measure the extent to which specified performance objectives are obtained. A range of performance standards is specified, on the basis of which student performance is differentiated. The practical difficulties in implementing criterion-referenced assessment are described. It must be considered whether the percentage increase in validity, reliability, and objectivity of grades awarded under a system of pure criterion-referenced assessment is likely to warrant the investment of teacher time, effort, and frustration that may be required. Three diagrams and five tables illustrate the discussion. (Contains 5 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 361 386

THE CONCEPT OF PERFORMANCE LEVELS IN CRITERION-REFERENCED ASSESSMENT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Mal Hewitson

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
MALCOLM HEWITSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

INTRODUCTION

This paper sets out to clarify the concept of performance levels in criterion-referenced assessment by applying the idea to the different types of tests commonly used in schools. These are mastery tests (including diagnostic tests) and achievement tests. Contract situations are also mentioned.

For explanatory purposes, assessment in terms of a single criterion is first discussed for each test type, even though in the case of many achievement tests multiple criteria are going to be used in real assessment situations.

In the course of discussing the basic concept of performance levels, practical issues are raised which imply that in some forms of achievement tests - essays, for example - the grading of students' performance levels is likely to continue to be influenced by norm-referenced thinking.

Statistical processes associated with determining cut-off scores or threshold standards, removing the effects of guessing, weighting particular criteria, scaling or rescaling scores, aggregating results, and so on are not discussed.

An appropriate starting point is the observation by Bourke et al (1981:7) that criterion-referenced assessment aims to determine the status of a student with respect to some well defined objectives. It refers to the grading of a student's work or performance on one or more

TM020486

 BEST COPY AVAILABLE

criteria. The criteria themselves are embodied in the stated objectives of the particular unit of work, for if tests are given in order to test the attainment of unit objectives, then clearly only criteria which are embodied in those objectives should be tested.

Under criterion-referenced assessment, Glaser and Nitko (Bourke et al 1981:7) suggest that levels of performance are interpreted in terms of specified performance standards. Such standards are carefully articulated as "models of achievement against which student performances are compared", to use Power's (1986:274) phrase. Clearly threshold standards (ie. the specified performance levels which will warrant the award of particular grades) need to take account of the students undertaking the unit of work, and may be adjusted upwards or downwards to accommodate this need. Whether they should be so adjusted (for example, in the case of certain basic minimal performance requirements) is, of course, another matter. In general, however, experience and expert judgement are the key factors in setting threshold standards, so that they are neither too easy nor too hard.

As suggested above, it is now proposed to explain the concept of levels of performance in mastery and achievement tests under a system of criterion-referenced assessment. Fundamental elements are examined in the first instance by reference to one criterion only, after which performance levels relating to multiple criteria are introduced.

PERFORMANCE LEVELS IN MASTERY TESTS

Mastery tests are defined here in a generic sense as tests which involve a single specified minimum standard - the threshold point or cut-off level - which must be attained for the student performance to

be considered acceptable. Such tests include those in which performance is judged on the following sorts of scales:

Fail/Pass

Unsatisfactory/Satisfactory

Not Competent/Competent

Non Mastery/Mastery

Conceptualised for One Criterion

A generalised concept of mastery tests involving only the one criterion, Knowledge of Content, may be represented as follows (Diagram 1):

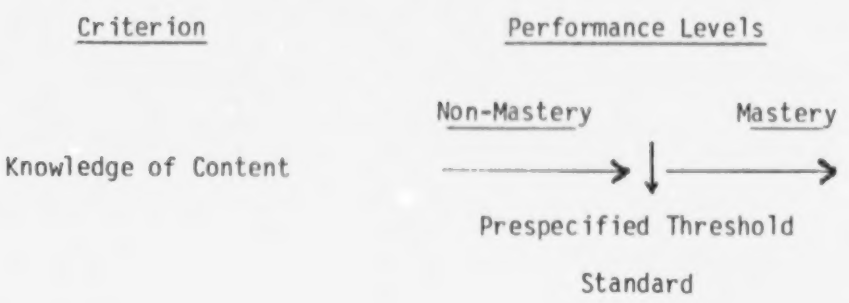

Diagram 1: Performance Levels in Mastery Tests - One Criterion

At this point, it is worth elaborating on a number of the elements contained in Diagram 1. First, the criterion, Knowledge of Content, is here defined to involve the factual recall of a well-defined domain of knowledge or subset of that domain. This criterion is presumed to have been embodied in the statement of objectives, one of which would state that students should know the concepts included in this unit of work. As noted earlier, if tests are given in order to test the attainment of unit objectives, only criteria which are embodied in the stated objectives should be tested.

Second, Diagram 1 theorises that non-masters may have some knowledge of content (but not enough to attain the threshold standard) and no upper limit is specified for relevant knowledge which masters may have. The diagram thus reflects the fact that the test is a sample of all items that could be asked on Knowledge of Content, and may not tap all the content knowledge of either non-masters or masters.

Third, it seems reasonable to propose that Power's (1986:275) comment with regard to Year 12 work, namely, that competence is continuous and gradually acquired rather than suddenly acquired by crossing a threshold, applies with equal force to the acquisition of competence in fundamental knowledge and skills. If so, the point is well taken since it suggests a certain arbitrariness in setting a threshold standard: in fact, what constitutes the prespecified standard is determined by teacher and/or expert judgement. This element of subjectivity enables some variation in establishing threshold standards, since some may consider, say, 75% correct responses as attaining content mastery, and others 80% or 85%.

In some cases, of course, all teachers may insist on 100% correct responses, as, for example, in safety training. Occasional exceptions, however, do not negate the fact of arbitrary judgements in setting threshold performance standards. As Bourke and Keeves (1977:36) have so succinctly put it, mastery does not imply perfection.

It is not only in setting standards, however, that the subjective judgement of teachers is involved but also in interpreting student performances. Apart from complications due to students guessing or giving a wrong answer when they actually know the correct answer, there may be occasions when teachers find student responses falling into grey areas; but they must nevertheless interpret such responses in terms of the specified threshold standards.

Though not apparent in Diagram 1, a fourth element of mastery tests is that they are usually taken by all students. This aspect is associated with the two general purposes of mastery tests, namely:

(i) The learning function, i.e. students need to attain threshold performance standards as an adequate foundation for further learning; and/or

(ii) The certification function, i.e. threshold performance standards need to be attained for certification purposes.

Conceptualised for a Number of Criteria

In any unit of work, it is common for a number of criteria to be generated by the formal statement of objectives. Assume for present purposes that five such criteria have been identified for a particular unit of work, as follows: Content, Comprehension, Application, Analysis and Evaluation.

For each criterion, a threshold performance standard is specified, the attainment of which signifies mastery according to teacher and/or expert judgement.

Diagrammatically, this position may be depicted as follows (Diagram 2):

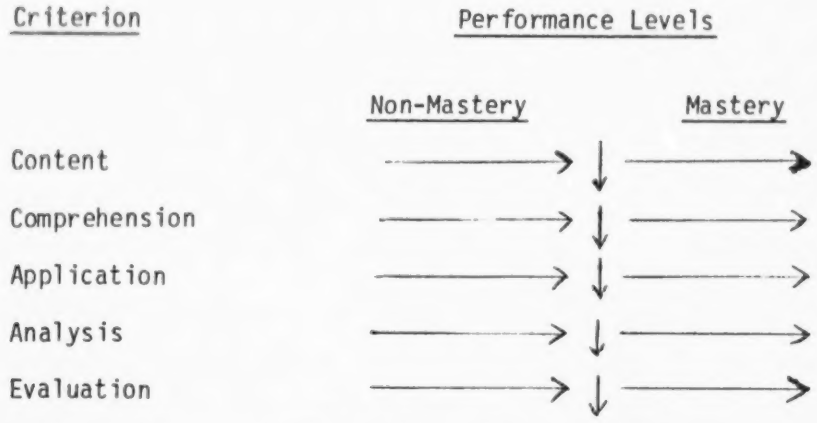

Diagram 2: Performance Levels in Mastery Tests - Multiple Criteria

In Diagram 2, a threshold standard is specified for each of five criteria. In theory at least, each criterion is treated as a discrete construct, and, therefore, each has its own scale, the bar " " representing the point at which the student is deemed to have attained mastery.

Suppose now that it was felt that the threshold standard for Content was too difficult. It would be a simple matter to lower the required standard making mastery easier to attain as far as students are concerned and reducing the importance of Content in the overall marking schema.

Similarly, if it was thought that it was too easy to attain mastery of, say, the Evaluation component, the threshold standard could be re-written at a more difficult performance level. This would have the effect of both making mastery harder for students to attain and increasing the importance of Evaluation in the overall marking schema. Thus, raising or lowering the threshold standard on any criterion is akin to weighting attainment on that criterion more heavily or more lightly. Hence it is inevitable that a form of weighting is integral to the present conceptualisation of performance levels, depending on where threshold standards are set.

A further consideration is that, although threshold performance standards are specified, it is possible to vary the requirements for what might constitute overall mastery of a unit of work. Such variation is achieved by changing the number and/or distribution of threshold standards which students must attain in order to be graded as masters of that unit of work. For example, it might be stated that students must attain mastery on all five criteria identified above in order to be graded as masters. Or perhaps mastery on four of the five criteria would be acceptable. Or, again, mastery on four criteria

provided that the four include Content and Application may be deemed acceptable. Clearly, the rules which are to govern whether or not students are graded as masters (referred to as a Decision Table) need to be stated at the beginning of the unit for all to see. As stated earlier, however, such grading is evidence that a specified performance level has been attained as required either for further learning or for certification purposes.

Conceptualised for Diagnostic Tests

Diagnostic tests are defined here as tests designed to reveal relevant weaknesses in a student's learning. Although test results may assist in the placement of students, the prime purpose is to identify student remediation needs.

The format used in diagnostic tests is that of mastery testing, in that relatively large clusters of items focus on relatively limited work segments, since detailed information is needed to detect learning deficiencies. As an example, Ahmann and Glock (1981:411) note that the Diagnostic Mathematics Inventory measures mastery of 325 behaviourally stated objectives in order to identify strengths and weaknesses across seven performance levels.

Conceptually, however, performance levels in diagnostic testing are merely a special category of performance levels in mastery tests discussed in earlier sections. Their focus, however, is the learning function rather than the certification function of such tests.

PERFORMANCE LEVELS IN ACHIEVEMENT TESTS

Conceptualised for One Criterion

Whereas mastery tests measure whether specified performance standards associated with stated objectives are attained or not, achievement tests are designed to measure the extent to which such objectives are attained. In achievement tests, therefore, under criterion referenced assessment a range of threshold performance standards is specified, on the basis of which student performances are differentiated between, say, levels 1, 2, 3, 4 and 5, as depicted in Diagram 3.

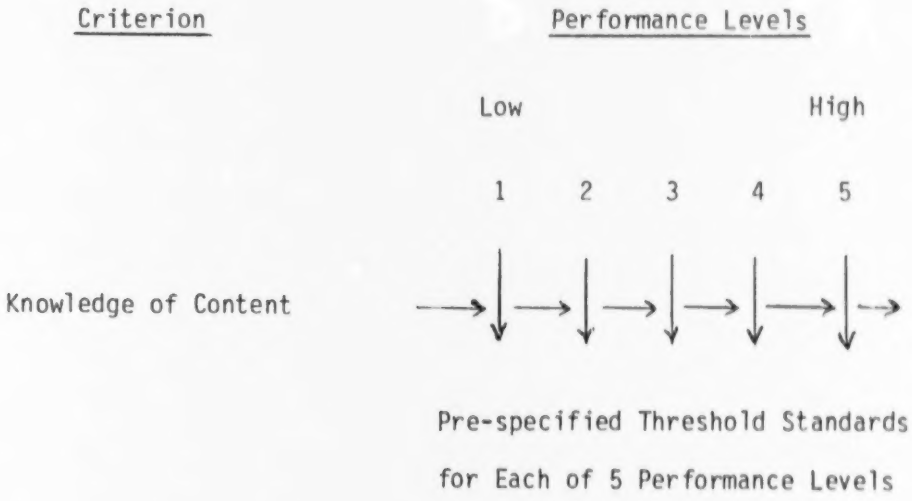

Diagram 3: Performance Levels in Achievement Tests: One Criterion

With regard to Diagram 3, there is again an assumption that the criterion, Knowledge of Content, is embodied in the statement of objectives. As with mastery testing, threshold standards are determined by teacher and/or expert judgement, with some arbitrariness resulting; and again, subjectivity plays a part in the interpretation of student performances in terms of threshold standards. Furthermore, although achievement tests are usually associated with larger domains

of knowledge than are mastery tests, as with mastery tests, the test material is a sample of all possible items that might have been asked. And finally, the tests are usually taken by all students in the group.

It is the differences between mastery and achievement tests, then, that assume conceptual significance. The first such difference is that performance standards are specified not for a single threshold standard but for a range of performances, typically from relatively low to relatively high, e.g. from relatively little content knowledge (Level 1) to relatively comprehensive content knowledge (Level 5) as depicted in Diagram 3.

A second difference is that, whereas the purposes of mastery tests are to ascertain whether students have adequate foundations for further learning and/or have earned certification that minimum performance standards have been attained, the purposes of achievement tests emphasise the following:

(i) motivating learners by rewarding higher achieving students with higher grades; and

(ii) certifying the quality of student achievement in terms of the full range of pre-specified levels of performance.

Two points need to be made at this juncture. The first is that a general purpose implicit in the two purposes of achievement tests articulated above is that of providing a basis for selection among students by other users of the assessment information. Thus, the information can be used to put students into streams, to select them for further education courses, to sort out applicants for jobs, and so forth. Mastery tests are not designed to provide such selection information.

The second point to be noted is that specification of different performance levels implies that performances can be compared in terms of quality, i.e., a Level 5 performance is better than a Level 4 performance, a Level 1 performance is not as good as a Level 2 performance, and so on. Such comparisons are typical of norm-referenced assessment, and are inevitable when test results are ranked for selection purposes.

Conceptualised for a Number of Criteria

In this section, the five criteria used earlier are again used to illustrate that there are two distinct conceptions of "levels of performance" under criterion-referenced assessment, depending on the way in which test results are interpreted and reported. The first is most readily understood and, probably, most commonly practised by teachers familiar with norm-referenced assessment. In this conception, once threshold performance standards are prespecified for each criterion established for testing, (in this example, Content, Comprehension, Application, Analysis and Evaluation), student performances are recorded in terms of the levels attained on each criterion and an averaged aggregate result for the test is awarded. For example, a student may attain Levels 3, 3, 4, 4, and 5 on the five criteria included on the test, and an averaged aggregate of Level 4 be awarded, thereby summarising performance on that test. Numerical equivalents for each level on each criterion may or may not be used in this process. Level 4, of course, might be a "C", or a "Pass" grade, or other result depending on the terminology used. Results from all tests are aggregated for final (end of term) reporting purposes, and overall weighting is achieved by giving more or less weight to particular tests in this aggregation process.

A different way of conceptualising performance levels for reporting achievement test results occurs when student performances are recorded in terms of the levels attained on each criterion but no attempt is made to obtain an aggregate result for any particular test. Instead, a record is kept of performance levels attained on each criterion on all tests administered for reporting purposes during the reporting period. At the end of that period (month, term, semester, etc.), attainment levels on each criterion are interpreted and, if necessary, aggregated to report an overall result for each student.

For purposes of clarification, a summary of test results extracted from a teacher's record book under this approach is illustrated below (Table 1). In this example, it is assumed that five performance levels have been specified for each criterion.

Table 1: Record of Student Test Performances

	Test 1		Test 2	
		Performance		Performance
Student	Criterion	Level	Criterion	Level
Joan Allen	Content	5	Content	4
	Comprehension	4	Analysis	4
	Application	4	Evaluation	3
Bill Bates	Content	3	Content	4
	Comprehension	3	Analysis	3
	Application	2	Evaluation	2

Given results such as those shown in Table 1, the Decision Table drawn up prior to the start of the unit of work would now be consulted in order to determine the final (end of term) grades to be awarded. A Decision Table contains the formal specification of the conditions

under which grades are to be awarded. Descriptions of each level of performance - what is required if the work is to be assessed as attaining each threshold standard on each criterion - are assumed to have been set down.

In the present example, the Decision Table (Table 2) reflects the fact that performance levels are aligned with grades in the following manner:

Table 2: Example of a Decision Table: Performance Levels Aligned with Grades

The following performance levels must be attained for an overall grade of:

FINAL GRADE	E	D	C	B	A
Content	1	2	3	4	5
Comprehension	1	2	3	4	5
Application	1	2	3	4	5
Analysis	1	2	3	4	5
Evaluation	1	2	3	4	5

There is, however, no a priori reason for the number of performance levels specified on any criterion to be the same as the number of grade levels which may be awarded. In other words, the fact that students may be awarded one of five grades does not automatically mean that five performance levels must be specified for each criterion, as in Table 2. A Decision Table in which different numbers of performance levels are specified for different criteria is depicted below (Table 3).

Table 3: Example of a Decision Table: Performance Levels not Aligned with Grades

The following performance levels must be attained for an overall grade of:

FINAL GRADE	E	D	C	B	A
Content	1	2	3	4	5
Comprehension	1	2	3	4	4
Application	1	2	2	3	4
Analysis	1	2	3	3	4
Evaluation	1	2	2	2	3

Table 3 shows that five levels of performance are specified for Content, four levels for Comprehension, Application and Analysis, and three levels for Evaluation. The Table also shows that a 4 level of performance on Comprehension is necessary for a B to be awarded overall, but note that the same performance level would qualify a student for the award of an A. On Application, a 2 level of performance could earn the student a D or a C. On Analysis, a 3 level of performance qualifies students for a C or a B, whereas on Evaluation, a 2 level of performance could earn a D, a C or a B. In general, Table 2 reflects the fact that specified performance levels are not necessarily aligned with particular grades.

A few moments of reflection reveal that the decision rules in Table 3 present a system of weighting. Weighting is achieved by specifying the threshold standards that must be attained in order to warrant the award of a particular grade. For example, on Comprehension, a 4 level of performance (the highest level specified) is needed for an overall B, indicating that this criterion is receiving a heavier weighting than say, Analysis, where a 3 level of performance (the second highest

level) could result in an overall award of a B. On the other hand, Evaluation is not weighted heavily since the same performance required for a D is deemed adequate for the award of a C or a B. Both the number of threshold standards and their difficulty level are subject to teacher and/or expert judgement.

Another aspect of this conception of criterion-referenced assessment and the interpretation of levels of performance addresses the matter of final grades. Strictly speaking, the overall grade awarded is determined by the weakest performance attained on any of the criteria tested, since the Decision Table specifies the minimum performance level that must be attained on each criterion to warrant the award of particular overall grades. For example, consider the following performance levels attained by Wendy Cool (Table 4):

Table 4: Record of Wendy Cool's Results

		Performance Level	Highest Grade to be awarded
Wendy Cool:	Content	5	A
	Comprehension	4	A
	Application	4	A
	Analysis	3	B
	Evaluation	3	A

According to the decision rules laid down in Table 3, Wendy qualifies for an A in Content, Comprehension, Application and Evaluation, but not in Analysis. As she attained only a 3 level of performance in Analysis, her overall grade is a B.

Though based on fundamental principles which underlie criterion-referenced assessment, Wendy's final result may seem rather

unfair: hence a perceived need for trade-offs to be introduced into the decision rules. Thus, in many assessment situations, teachers may allow very good performances on some criterion to build up credit, so to speak, to offset weaker performances on another criterion. For example, some teachers may feel that Wendy Cool's work merits an A overall. If so, the relevant decision rules would need to be incorporated into the pre-specified Decision Table.

It should be noted that such trade-offs compromise the logic of criterion-referenced assessment. This is so because a founding principle of the criterion-referenced approach is that the criteria need to be defined as discrete qualities: if they are not, confusion arises on two counts: first, it is no longer certain what it is that has been awarded a grade; and second, particular aspects of a students' work may contribute to two or more overlapping criteria, and thus be double-counted.

If it is assumed that, in addition to being discrete qualities, the criteria being used for assessment are not trivial, then the threshold performance standard should be attained on all criteria - not four out of five - for a student to be awarded an overall grade which reflects the defined level of achievement. This is why purists would insist that, under criterion-referenced assessment, a student's overall grade cannot be higher than the lowest threshold standard attained. The nigger in the woodpile, of course, is the socially-based need to aggregate grades awarded on conceptually discrete criteria into a single index of overall achievement; and there seems to be little likelihood of that need diminishing in the near future.

It is worth pursuing the notion of trade-offs a little further because they may be used to resolve the "breadth vs depth" problem frequently faced by teachers when marking certain items of assessment. In the

case of essays, for example, credit built up for breadth of topic coverage may be used to compensate for some lack of depth, or fluency and style may raise the higher overall performance level attained and the grade awarded.

In passing, a potential weakness of criterion-referenced assessment lies in the fact that this approach appears to demand a trait by trait analysis of essay work, since grades are to be awarded for each of the criteria applicable to the particular essay task. There is no agreement in the literature, however, that analytic assessment of essay work is as valid and reliable as the holistic approach, in which a global impression has most influence in assessing a piece of writing. In analytic assessment, Cooper (1984:34) suggests that the essay should be read rating one trait at a time to keep impressions of one trait from influencing the rating on another. Otherwise validity and reliability may be lost. Hence analytic rating becomes time-consuming and tedious, and, according to Quellmalz (in Cooper 1984), can take four, five or six times longer than a holistic marking of the same paper. Furthermore, there may be a tendency for raters to focus unduly on weaknesses as they read the essay concentrating on a single trait. Finally, Cooper (1984:36) points out that, using analytic marking, the best or most noteworthy essay frequently does not achieve the highest grade as the analytic approach may fail to reflect the distinctive value of the composition of the essay. Lloyd-Jones (in Cooper 1984) agrees that the whole is more than the sum of the parts, but notwithstanding that, feels that the categorisable parts may be too numerous and too complexly related to permit a valid report based on trait by trait analysis. Thus, it may be that, marking some assessment items, the logic of criterion-referenced assessment <u>should</u> be compromised on educational grounds.

Certainly the discussion initiated by Wendy Cool's results reveals that there can be real difficulties surrounding the implementation of criterion-referenced assessment with its associated student performance levels. Another difficulty is revealed by the results attained by one of Wendy's classmates, Don David, who has attained the following results (Table 5):

Table 5: Record of Don David's Results

		Performance Level		Highest Grade to be awarded
		Test 1	Test 2	
Don David:	Content	3	4	C (Test 1) B (Test 2)
	Comprehension	4	-	A
	Application	2	-	C
	Analysis	-	3	B
	Evaluation	-	2	B

The decision rules (refer to Table 3) indicate that Don's work is to be awarded a C overall.

The problem raised by Don's results centre on the fact that his performance level on Content was 3 in the first test and 4 in the second. What rating, then, is his overall attainment on Content to be awarded? Some way of dealing with such problems needs to be articulated, whether the content itself is the same for both tests or different for each test. A very poor performance on one criterion tested early in the term might otherwise condemn a student to an unfairly low overall term grade in that subject, given strong improvement on the part of the student during the term. Again, the

need is to specify in advance the decision rules that are to apply so that all interested parties are fully aware of them.

There is little to be gained by elaborating the details that need to be recorded by teachers putting the conception of criterion-referenced assessment and performance levels into practice. Power (1987:279) has estimated that a Year 12 student taking six subjects with six descriptions per subject could end up with a report seven or eight pages long. Docking's (1987:12-15) paper also gives an idea of the extensive recording and grading process required of the teacher.

In this regard it is important to consider again the purposes of the assessment. The fine details may be of use to teachers and provide feedback on progress to students, but, for screening or selection purposes, highly detailed reports are of little value to external users of the assessment information. They are likely to seek some aggregation of minutiae into overall grades (or other index) which summarise each student's level of performance. However, as indicated above, aggregation of performance levels across criteria contradicts a basic premise of criterion-referenced assessment that criteria should be conceptually distinct from each other, which means that conceptually speaking, performance levels on different criteria cannot be added together to give a meaningful summary grade. However, once a range of aggregate grades is awarded, performance levels can inevitably be compared with each other in that an A represents a better performance than a B, and so on, just as in the case of norm-referenced assessment. All achievement tests, irrespective of whether they are interpreted with reference to norms or criteria, are designed to spread student results across a range of achievement levels, thus making comparisons of performance possible.

PERFORMANCE LEVELS IN CONTRACT SITUATIONS

The importance of identifying the purposes of assessment becomes apparent when consideration is given to specifying different levels of competence usually associated with achievement tests. An important function of achievement tests is to differentiate between student performance levels for selection purposes. However, there are circumstances in which teachers may specify performance standards the attainment of which is to be recognised by the award of the highest grades. The practice involves contracting between teachers and high aspiring students. For example, a teacher may specify a set of superior performance standards attainment of which would result in the award of, say, a Distinction.

Such an approach is like mastery testing in that the threshold standard is either attained by students or not attained.

However, there are also significant differences, in particular, with regard to the purposes of the test. Thus, the learning function in the present example has to do with challenging higher achieving students, and the certification function includes the identification of such students for possible selection purposes. A main difference between this sort of contracting and commonly used achievement tests (which specify a range of attainment levels) is that in the present case there is an expectation that only higher achieving students will try for the superior standards of performance.

Finally, an alternative to specifying higher performance standards on existing criteria is to introduce a new element into the test (for example, a task involving an evaluation) in order to challenge and extend the performance of high aspiring students.

CONCLUDING COMMENTS

Having examined the concept of performance levels in criterion referenced assessment, by way of conclusion it seems appropriate to make a few related points of a general nature.

The first comment has to do with the fact that while it is conceptually possible to separate out discrete criteria, to specify threshold standards and to interpret student performances in terms of those criteria and standards, in fact all three processes are subject to challenge: complexity rather than simplicity characterises criteria; subjectivity rather than objectivity is fundamental to the setting of threshold standards within the reach of and meaningful for students; and judgemental interpretations of student performances are notoriously fallible given the present state of test construction, administration and marking. In the case of project reports and essay work in particular, criterion-referenced grading may actually be educationally counterproductive. Furthermore, when marking a batch of essays, in practice teachers may find it impossible not to compare the essay now being graded to the essay graded a few minutes ago, an unacceptable state of affairs in criterion-referenced assessment.

Another practical difficulty may arise if the grading and recording of student performance become fragmentary, burdensome or time consuming: there is more chance of increasing the incidence of mechanical errors of transcription.

To these notions can be added the need to aggregate grades awarded on discrete criteria into a single index and perhaps produce an order of merit for purposes of selection (which is a norm-referenced requirement). Certainly the practicalities associated with the selection function of assessment cannot be ignored by the schools.

The question that must be asked, then, is whether the percentage increase in validity, reliability and objectivity of the grades awarded under a system of "pure" criterion-referenced assessment is likely to warrant further significant investments of teacher time, effort and frustration which may be required to achieve it.

REFERENCES

Ahmann, J.S. and Glock, M.D. (1981) Evaluating Student Progress, Sixth Edition. Boston: Allyn and Bacon.

Bourke, S.F., Mills, J.M., Stanyon, J. and Holzer, F. Performance in Literacy and Numeracy: 1980. Melbourne: Australian Education Council.

Cooper, P.L. The Assessment of Writing Ability: A Review of Research. Princeton, N.J.: Educational Testing Service.

Docking, R.A. (1987) "Constructive Credentialling" in Unicorn, 13:1, pp. 10-17.

Power, C. (1987) "Criterion-Based Assessment, Grading and Reporting at Year 12 Level" in Australian Journal of Education, 30:3, pp. 266-284.